

#### **Barcode of Life**

Banbury Conference Center, Cold Spring Harbor, Feb. 2003

**Sponsored by COML** 

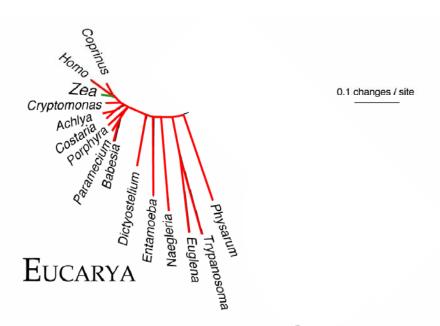


Mitochondrial encoded Cytochrome Oxidase A molecular metric for relating all genera and species.

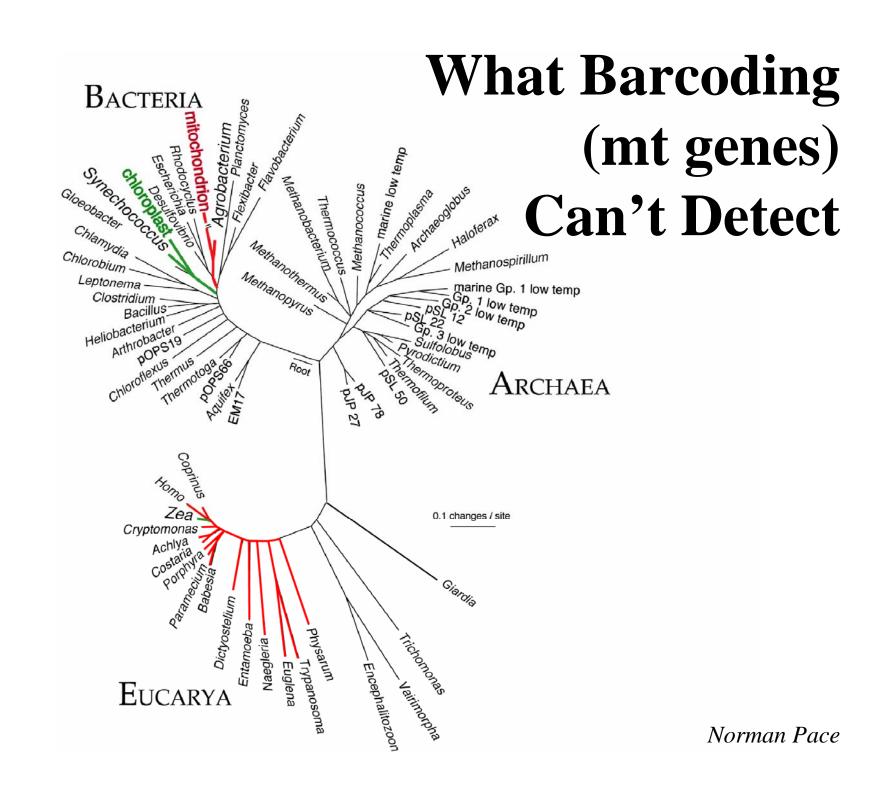


Tricorder
Any resemblance
purely coincidental.

# What Barcoding (mt genes) Can Detect



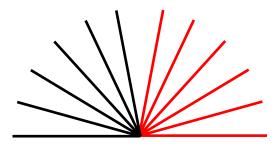
Norman Pace



#### Relative abundance and productivity of Marine Life

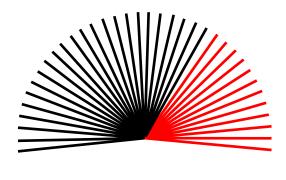
	Size B	iomass	Primary Production	Secondary Production
Prokaryotes	< 3 <u>μm</u>	50-82%	91%	
<b>Protists</b>	< 0.3mm	18.0%	9%	
Zooplanktor	1 < 3 cm	0.30%		93.0%
Swimmers	< 3m	0.07%		7.0%
Megafauna	> 3m	0.01%		0.5%
Millions Tor	ns	145,000	50,000	7,400

#### **Bacterial Diversity**



1987

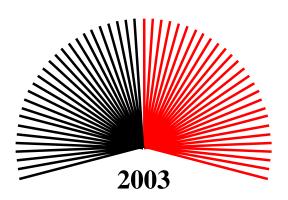
12 divisions: 12 cultured/ 0 candidate



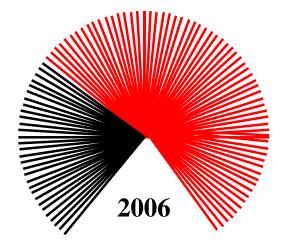
**1997** 

36 divisions: 24 cultured/ 12 candidate





53 divisions: 26 cultured/27 candidate



~100 divisions: 30 cultured/~70 candidate



The Census of Marine Life is a growing global network of researchers in more than 70 nations engaged in a ten-year initiative to assess and explain the diversity, distribution, and abundance of marine life in the oceans -- past, present, and future.

#### **International Census of Marine Microbes**

Goal: To report what is Known, what is Unknown but knowable, and what may be Unknowable about the diversity of marine microorganisms.

Determine the range of genetic diversity and relative numbers of different microbial organisms at sampling sites throughout the world's oceans

#### http://icomm.mbl.edu

PIs M.L. Sogin at MBL and J.W. de Leeuw at NIOZ Supported by the Alfred P. Sloan Foundation



#### **Questions that Drive ICoMM:**

How does diversity relate to function and ecosystem processes?

How does the choice of gene influence diversity assessments and inference about presence or absence of functional groups in a complex community?

What scales of heterogeneity – spatial, temporal are most appropriate for the census?

# After four meetings of four working groups and the SAC What Have We Learned?

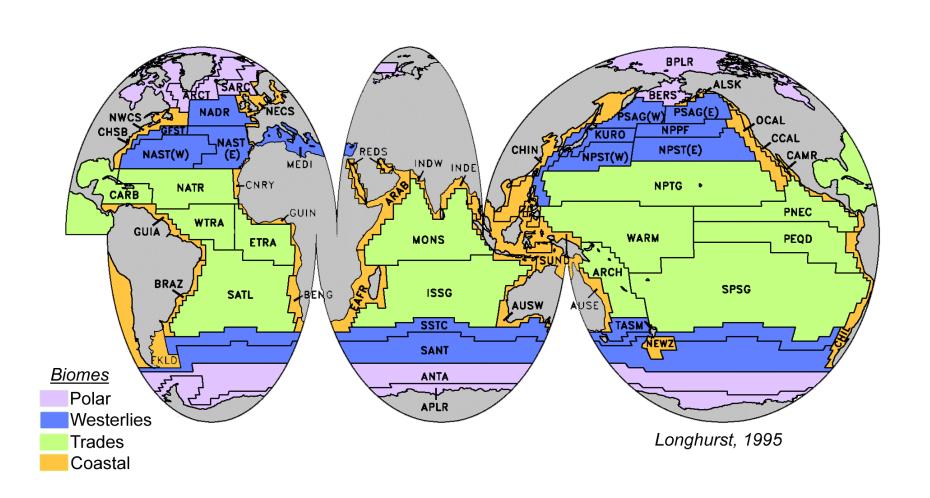
- 1. Information about community composition and relative numbers of different kinds of organisms is of key importance to most microbial oceanography investigations.
- 2. The metric is molecular; sequences, lipids, possibly proteins.
- 3. Contextual information is essential for meaningful interpretations.
- 4. The dynamic nature of the marine environment requires temporal sampling and therefore greater resource demands.
- 5. Sampling scales range from sub-millimeter to kilometer.
- 6. The Census is a big job; even in its most simple form, conventional technology will be insufficient!

Today's typical microbial surveys (sequences from 500-1000 amplicons) capture only a fraction of the population structure.

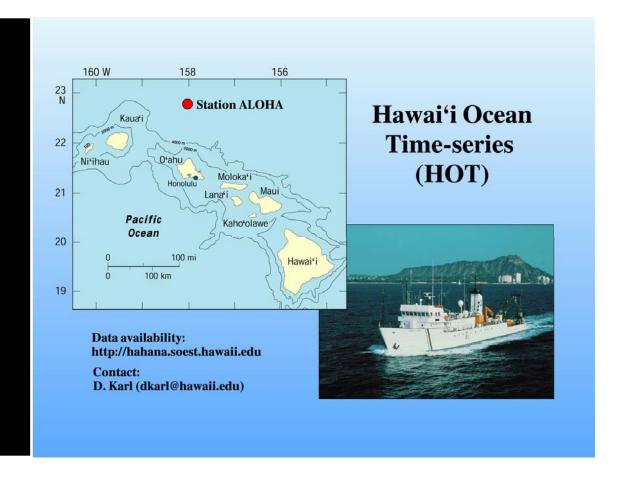
# After four meetings of four working groups and the SAC What Have We Learned?

- 1. Information about community composition and relative numbers of different kinds of organisms is of key importance to most microbial oceanography investigations.
- 2. The metric is molecular; sequences, lipids, possibly proteins.
- 3. Contextual information is essential for meaningful interpretations.
- 4. The dynamic nature of the marine environment requires temporal sampling and therefore greater resource demands.
- 5. Sampling scales range from sub-millimeter to kilometer.
- 6. The Census is a big job; even in its most simple form, conventional technology will be insufficient!
- 7. There was no consensus about optimal sampling strategy!

#### Biogeochemical provinces: a basis for examining geographic variation



- Established in 1988 as part of the U.S. JGOFS program
- Primary objectives: characterize timedependent dynamics in carbon, nitrogen, and phosphorus inventories and fluxes.



#### **Questions that Drive ICoMM:**

How does diversity relate to function and ecosystem processes?

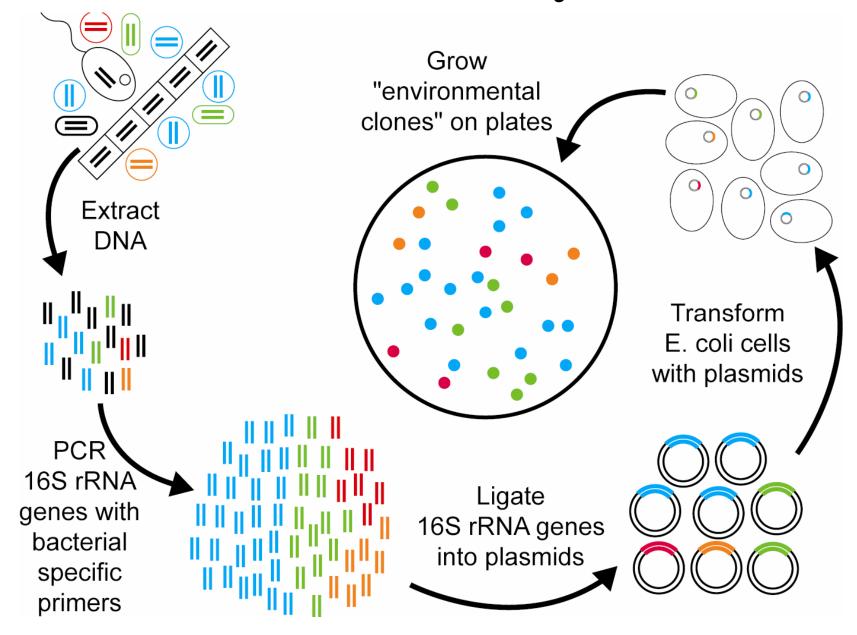
How does the choice of gene influence diversity assessments and inference about presence or absence of functional groups in a complex community?

What scales of heterogeneity – spatial, temporal are most appropriate for the census?

How can we link diversity at different scales?

What is the optimal measure of microbial diversity?

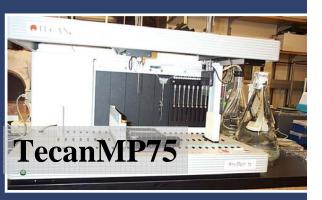
## **Traditional Clone Library Construction**



## Instrumentation

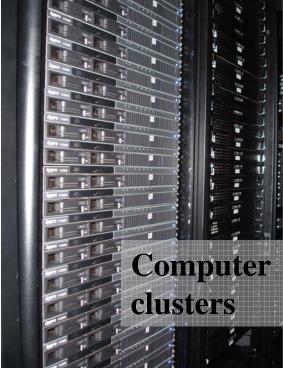














#### **Challenge of monitoring Bacterial populations**

- 1. Sequencing costs (~\$1/read) constrain the size of molecular surveys.

  Today's typical microbial surveys (sequences from 500-1000 amplicons) capture only a fraction of the population structure.

  Difficult to detect under-represented members of microbial communities
- 2. Communities dominated by a few abundant taxa will mask appearance of rare community members.
- 3. Surveys of a few hundred rRNA genes do not fully describe a microbial community.

Populations represented in most molecular surveys

CarlosPedro's-Alio TRENDS in Microbiology Vol.14 No.6 June 2006

#### **Challenge of monitoring Bacterial populations**

- 1. Sequencing costs (~\$1/read) constrain the size of molecular surveys.

  Today's typical microbial surveys (sequences from 500-1000 amplicons) capture only a fraction of the population structure.

  Difficult to detect under-represented members of microbial communities
- 2. Communities dominated by a few abundant taxa will mask appearance of rare community members.
- 3. Surveys of a few hundred rRNA genes do not fully describe a microbial community.
- 4. Information required about both diversity and relative abundance of different kinds of microbes.
- 5. DNA Arrays Only detect what is spotted or printed on array.

#### **Challenge of monitoring Bacterial populations**

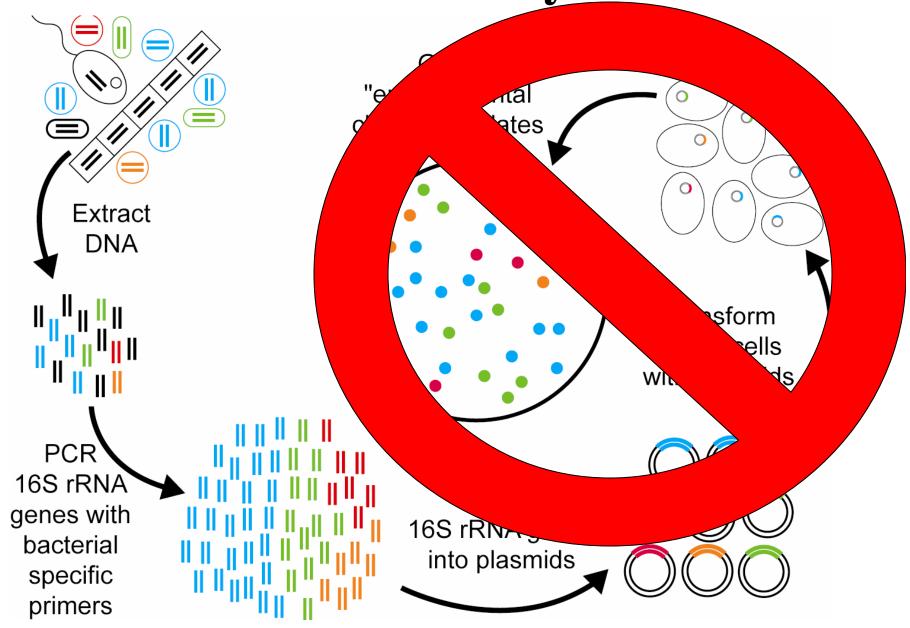
- 1. Sequencing costs (~\$1/read) constrain the size of molecular surveys.

  Today's typical microbial surveys (sequences from 500-1000 amplicons) capture only a fraction of the population structure.

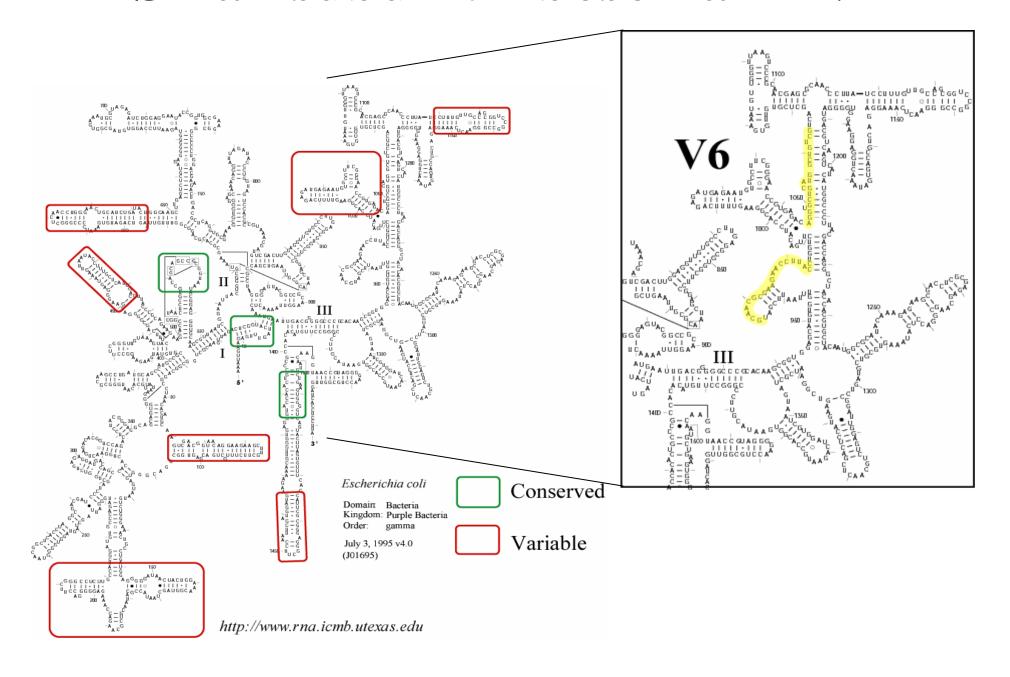
  Difficult to detect under-represented members of microbial communities
- 2. Communities dominated by a few abundant taxa will mask appearance of rare community members.
- 3. Surveys of a few hundred rRNA genes do not fully describe a microbial community.
- 4. Information required about both diversity and relative abundance of different kinds of microbes.
- 5. DNA Arrays Only detect what is spotted or printed on array.

TAG sequencing of hypervariable regions in rRNAs - a potential solution

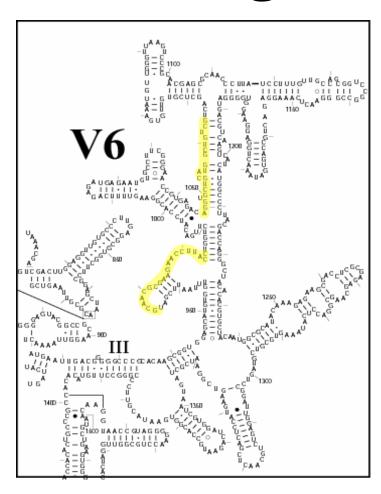
Traditional Clone Library Construction



## Small subunit ribosomal RNA



## V6 region of the SSU rRNA



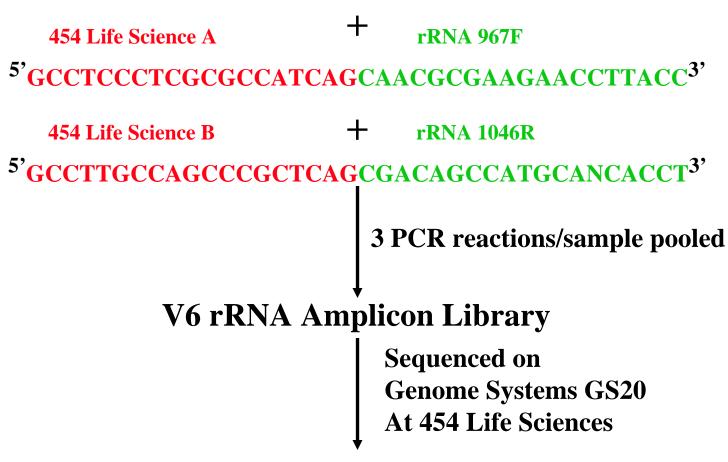
- Sequence many V6 regions
- Each serves as proxy for microbe in a community
- Use each V6-Tag sequence to query a reference data base
- Identify putative taxonomic source of each V6-Tag to infer community composition.

Provides an estimate of both diversity and the relative numbers of different kinds of microbes in a community - but it is

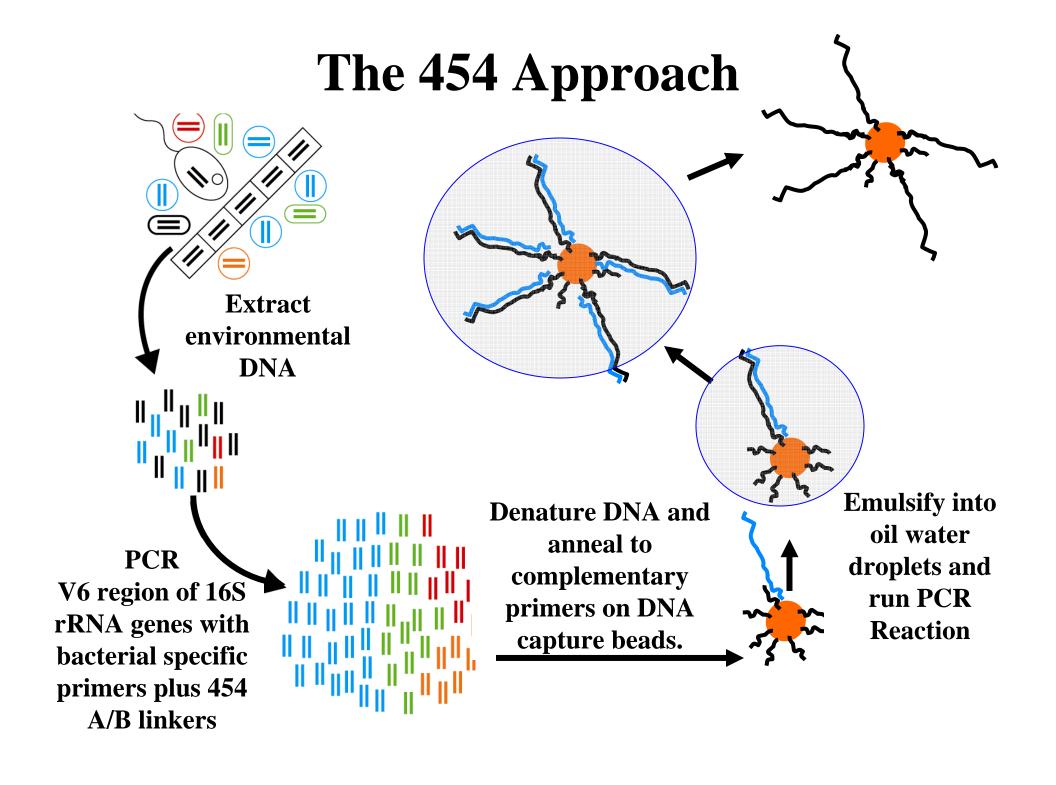
NOT A PHYLOGENETIC TOOL!

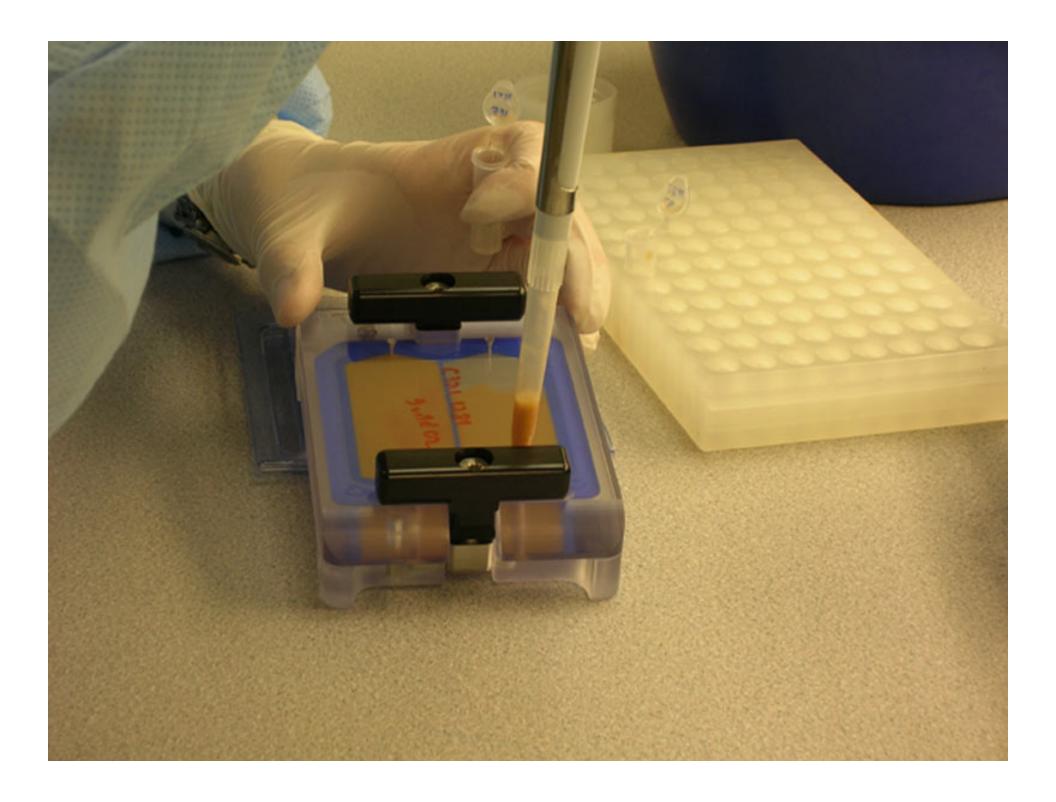
## The 454 "Tag" Approach

#### **Environmental DNA**

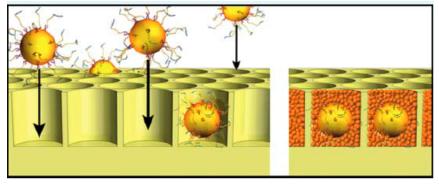


V6 rRNA Sequence Tags (6,000-25,000 tags/sample)

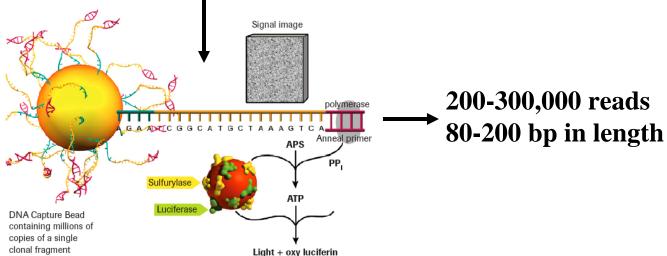




## The 454 "Tag" Approach



Add beads to the DNA Bead incubation mix (containing DNA polymerase) in PicoTiterPlate<sup>tm</sup>device - 1 bead/well

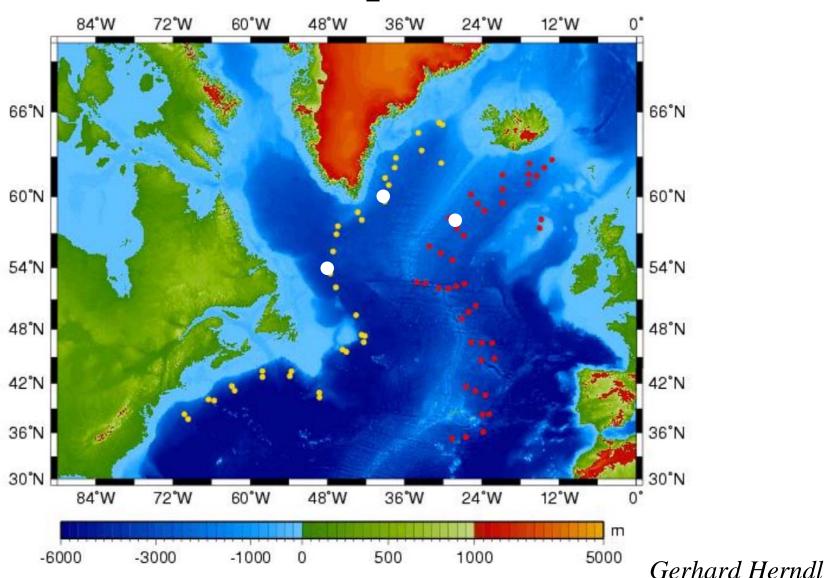


Flow sequencing reagents (containing buffers and nucleotides) across the wells of the plate and perform solid phase pyrosequencing.

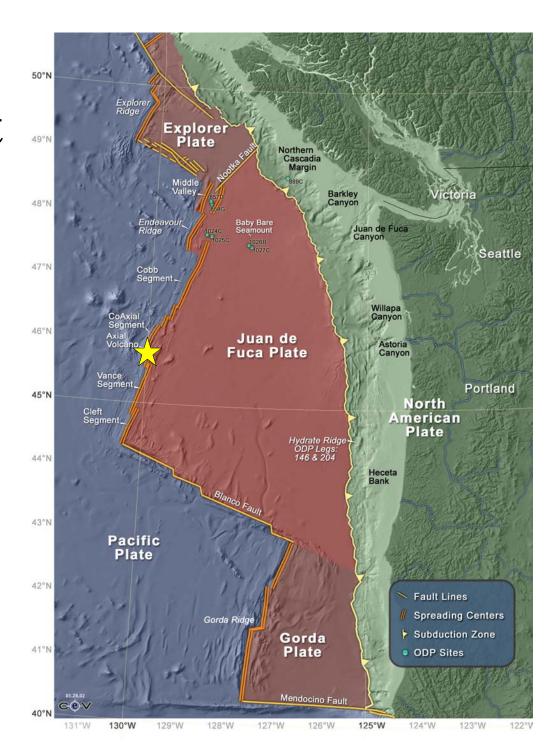
## Identifying the Known Universe: Creating a V6 Reference Dataset

- Collected full-length 16s sequences (containing V6) from four major sources: NCBI, ARB, LBL, and RDP
- ~ 120,000 unique full-length 16s sequences- taxonomy from RDP II Classifier
- 44,000 unique V6 sequences
- All full-length and V6 sequences are included in both a BLAST-able format and a searchable SQL database

# TRANSAT Samples North Atlantic Deep Water (NADW)

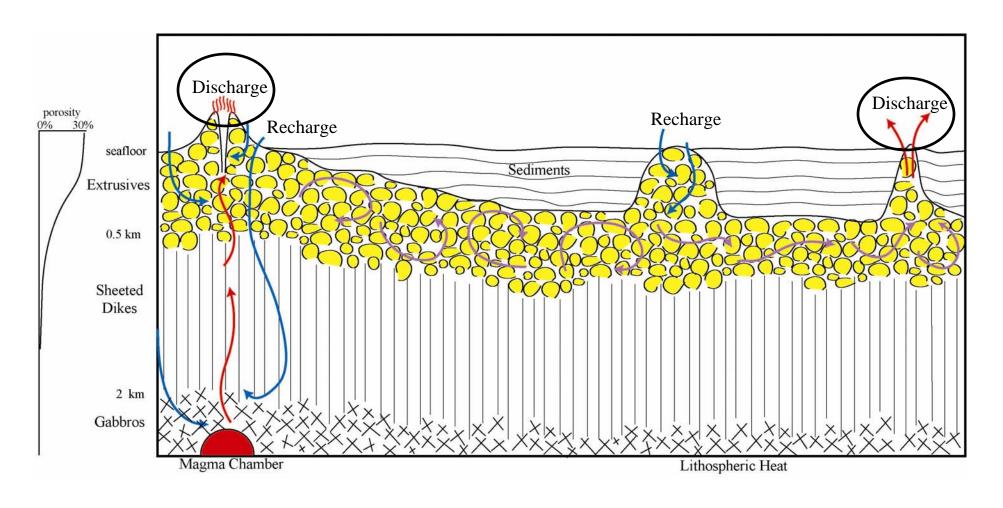


## Axial Seamount



## The crustal aquifer as a microbial habitat:

Linking deep seawater, sediments, and crustal fluids



Decreasing porosity and permeability



- Remove low quality data
- •Trim 5' primer sequence
- •Trim 3' sequence to end of V6 > ~60 bp

BLAST against 44,000 V6sequence database

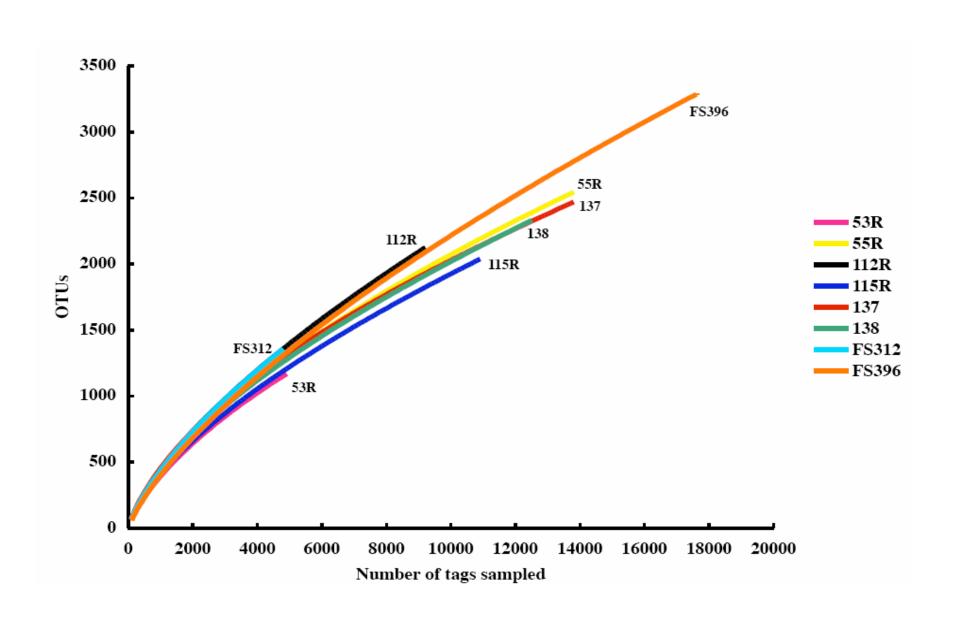
Align query with best scoring V-6 sequences

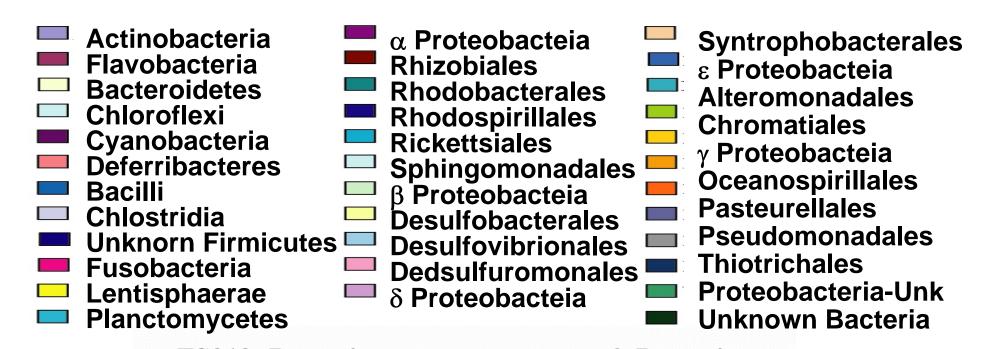
Minimum distance identifies closest entry in reference DB.

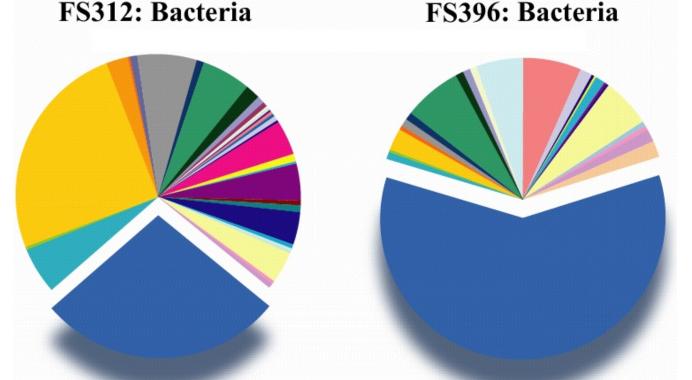
Rarefaction Analysis

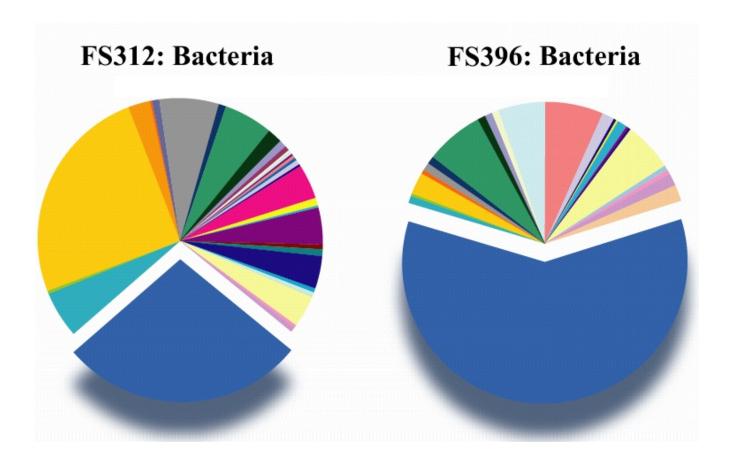
Phylotype distributions

### Rarefaction: BLAST OTUs







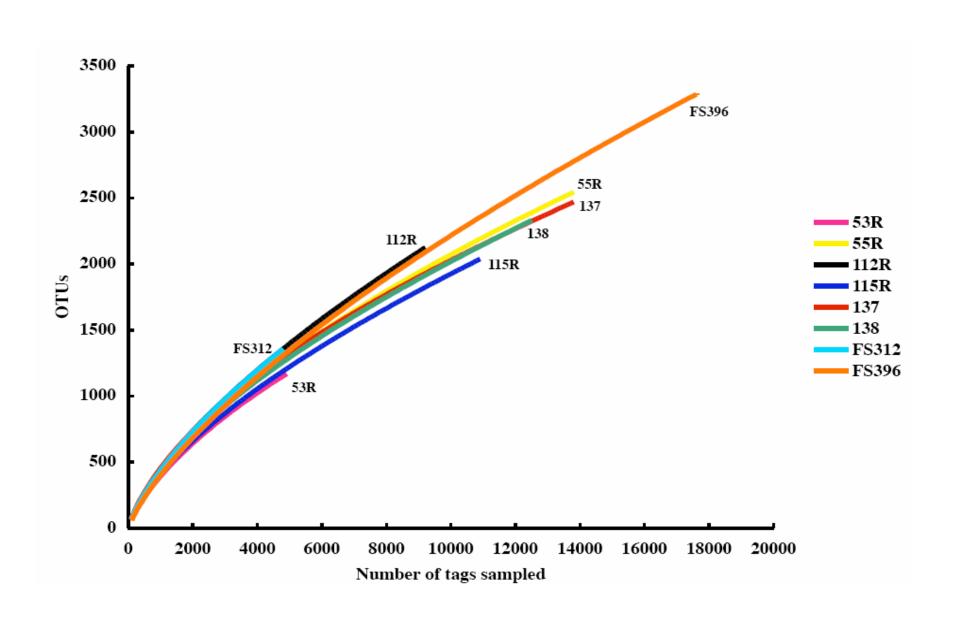


7.2	$H_2S/\Delta T$ (μmol kg <sup>-1</sup> °C <sup>-1</sup> )	18.9
6.26	pН	5.08
2.4	Alkalinity (mEq/L)	3.7
19.8	Mn (μmol/kg)	4.8
0.8	Fe (µmol/kg)	7.9

# Data Summary

Sample ID	Total Reads	Trimmed Tags	Unique Tags	BLAST OTUs
53R	6,505	5,000	2,656	1,184
55R	18,439	13,902	7,187	2,555
112R	12,916	9,282	5,752	2,135
115R	14,731	11,005	5,777	2,049
137	18,137	13,907	6,752	2,480
138	18,451	14,374	7,168	2,550
FS312	6,605	4,835	2,769	1,362
FS396	22,994	17,666	8,699	3,290

## Rarefaction: BLAST OTUs



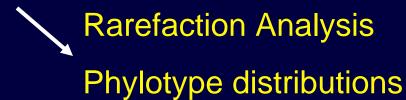
- Remove low quality data
- •Trim 5' primer sequence
- •Trim 3' sequence to end of V6 > ~60 bp

BLAST against 44,000 V6sequence database

Align query with best scoring V-6 sequences

Minimum distance identifies closest entry in reference DB.

**DOTUR** cluster analysis



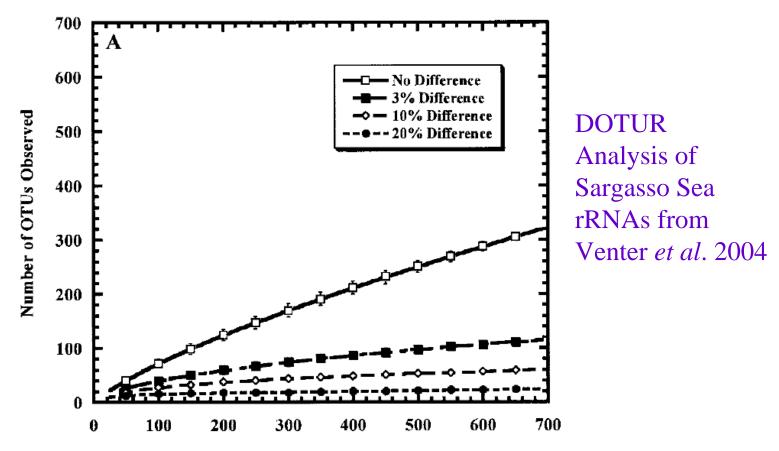
# Introducing DOTUR a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness

Patrick D. Schloss and Jo Handelsman\*

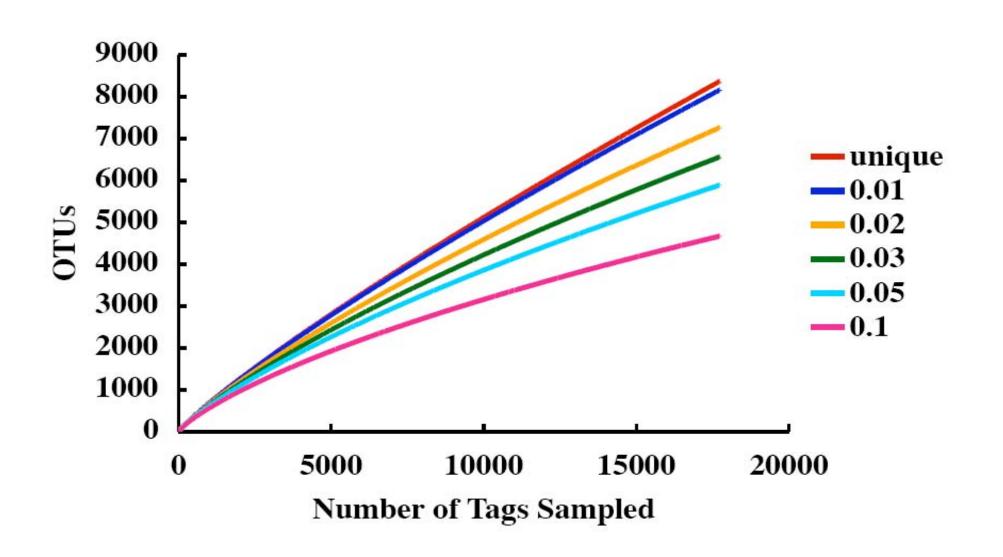
- •DOTUR assigns sequences to operational taxonomic units (OTUs) based on genetic distances between sequences
  - •Uses either the furthest, average or nearest neighbor algorithm for each distance level
- •The frequency at which each OTU is observed is used to construct rarefaction and collector's curves for various measures of richness and diversity

# Introducing DOTUR a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness

Patrick D. Schloss and Jo Handelsman\*



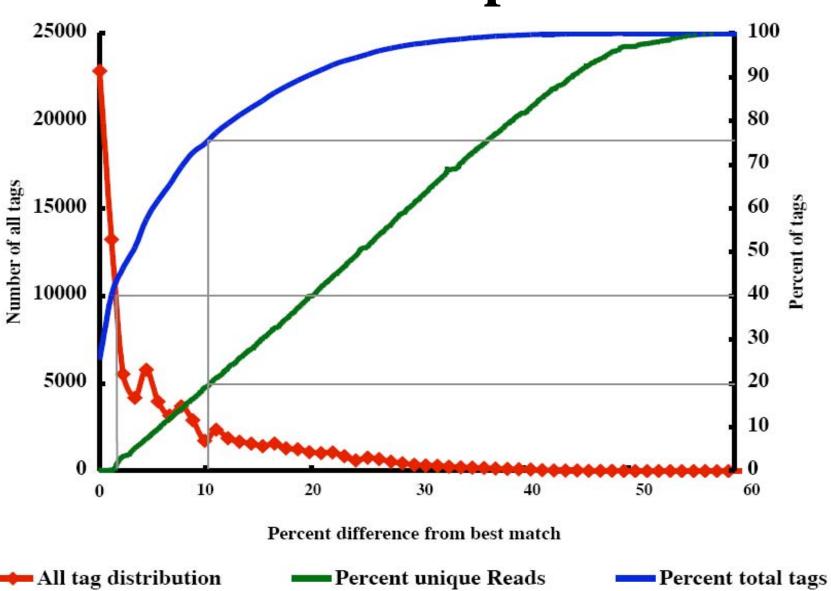
## Rarefaction: DOTUR FS396



### **DOTUR Analysis of 454 Sequence Tags**

Dist			0.0			0.01			0.03	
ID	Reads	OTU	ACE	Chao1	OTU	ACE	Chao1	OTU	ACE	Chao1
53R	4,995	2,595	16,488	14,565	2,527	14,550	12,993	1,954	6,826	6,271
55R	13,906	7,006	46,134	40,638	6,814	39,941	35,099	5,480	21,051	19,252
112R	9,270	5,459	36,213	31,948	5,272	31,276	27,388	4,238	15,620	13,223
115R	10,987	5,595	37,944	32,428	5,415	31,551	26,311	4,209	14,276	12,499
137	13,898	6,594	41,117	36,007	6,409	36,022	31,557	4,782	15,707	13,782
138	14,375	7,055	51,102	43,875	6,891	44,438	38,786	5,666	25,627	23,342
FS312	4,849	2,661	15,020	12,905	2,613	13,386	11,315	1,968	5,910	5,676
FS396	17,731	8,376	59,231	49,964	8,169	51,611	42,933	6,568	25,908	23,081

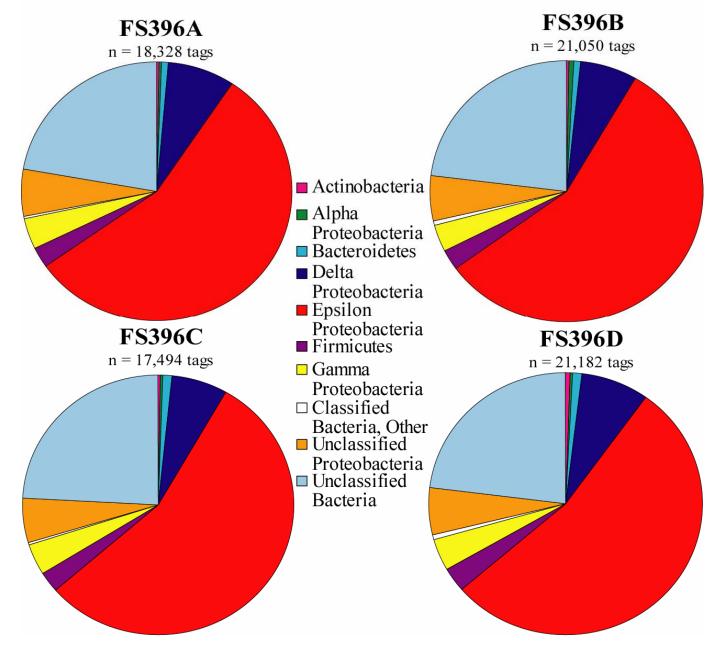
# Rare Biosphere



# Concept of Rare Biosphere

- Minor population members may serve as keystone species
- May simply reflect biogeography and many yet to be discovered habitats
- Rare Biosphere may be an important source of genomic innovation being observed in new microbial genome sequences
- Under-represented microbial populations may become dominant in response to environmental shifts
- Rare biosphere may serve as sentinel for global change

## Reproducibility Results: FS396



V	VAMPS-Visualization and Analysis of Microbial Population/Structures									
	HEATMAP									
	<ul> <li>⊙ RELATIVE DISTRIBUTION</li> <li>⊙ ABSOLUTE DISTRIBUTION</li> <li>⊙ TAXONOMY WITHIN SITE</li> <li>⊙ SITE WITHIN TAXONOMY</li> <li>⊙ DUMP HEATMAP</li> </ul>									
CO	MP	ARISON CH	ART TYPE DISTANCE METRIC	TAXONOMIC RANK	OTU %SIMILARITY					
HORIZONTAL B VERTICAL BAR PIE CHART by V EUCLIDE PIE CHART by T CANBER			MANHA EUCLIDE	PHYLUM CLASS ORDERX	8() 8! 9( 9(‡					
			GEO-SPATIAL	AD-HOC SEARCH	METHODS					
			ESTIMATION METHOD ENVIRONMENTAL DATA	SEARCH METHOD	STATISTICAL METHODS					
			AS IS NONE	BY TAXO	MANTEL TEST					
			GARP SST	BY OTU	LINEAR REGRESSI					
BBR CHLOROPH BY TAG SONS KRIG SALINITY SONS										
			IXIO T	DI TERSOY	NON-PARAMETR					
DAT	AS	ETS								
All	All All Project Datasets									
		Chicken Gut	□ STF7							
		Eel Pond	□ winter							
		Microbial Diversity	□ FS312_1000bp □ FS312_100bp □ FS312_400bp □ FS396_1000bp □ FS396_100bp □ FS396_400bp							
		Mount Hope Bay	☐ ST1S_021606F ☐ ST1S_081605F ☐ ST2S_021606F ☐ ST2S_081605F ☐ ST5S_021606F ☐ ST5S_081605F ☐ ST7S_021606F ☐ ST7S_081605F							
		Mouse Gut	□ C46G1 □ R43S2 □ R45G1 □ T43S1							
		North Atlantic Deep Water	□ 112R □ 115R □ 137 □ 138 □ 53R □ 55R							
		Sand	□ HS122       □ HS123       □ HS124       □ HS125       □ HS126       □ HS127       □ HS129       □ HS130       □ HS132       □ HS133       □ HS134       □ HS135       □ HS136       □ HS137       □ HS138         □ HS139							
		CTDBTL12								

FS521 🗎 LOIHI-CTD03 🗎 LOIHI-PP1 🗎 LOIHI-PP2 🗎 LOIHI-PP4 🗎 LOIHI-PP5 🗎 LOIHI-PP6

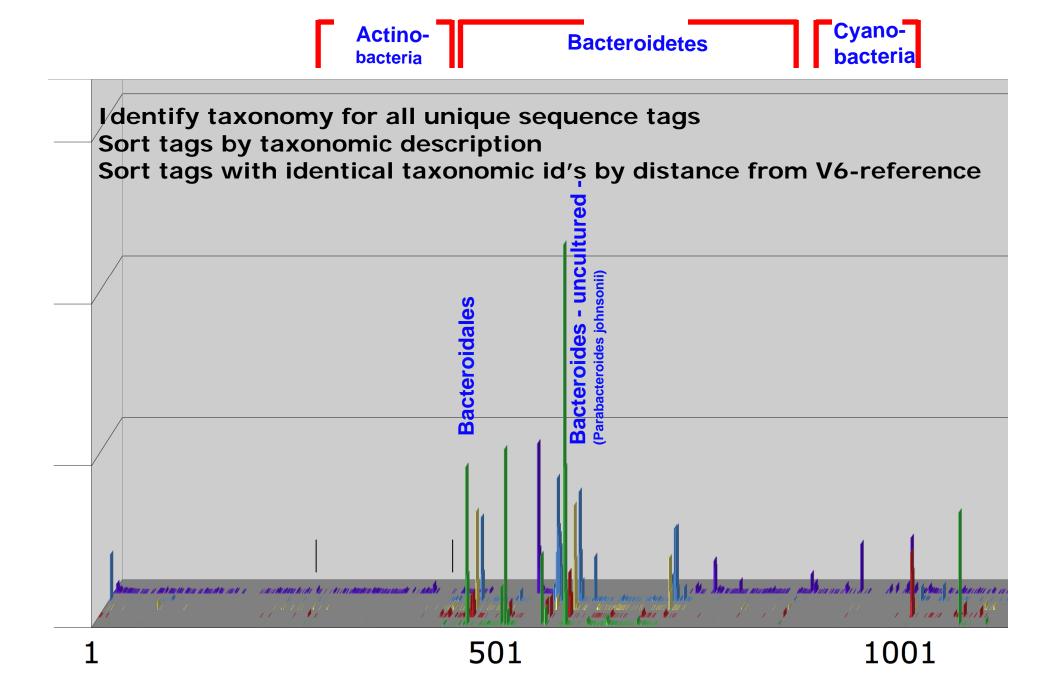


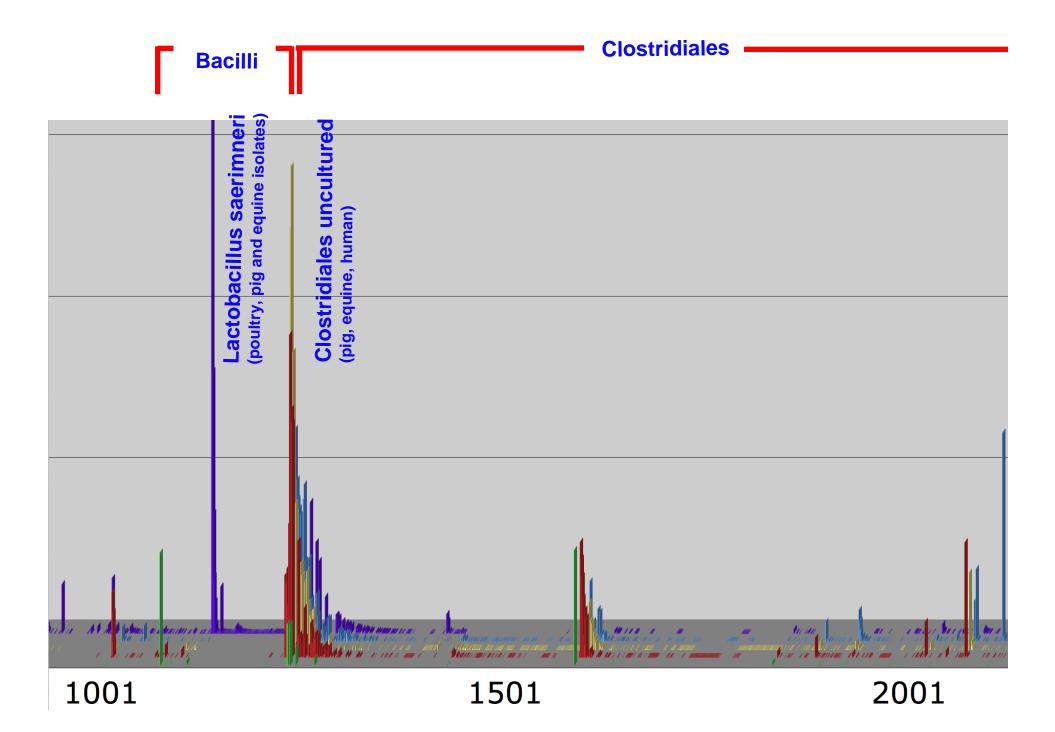
QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

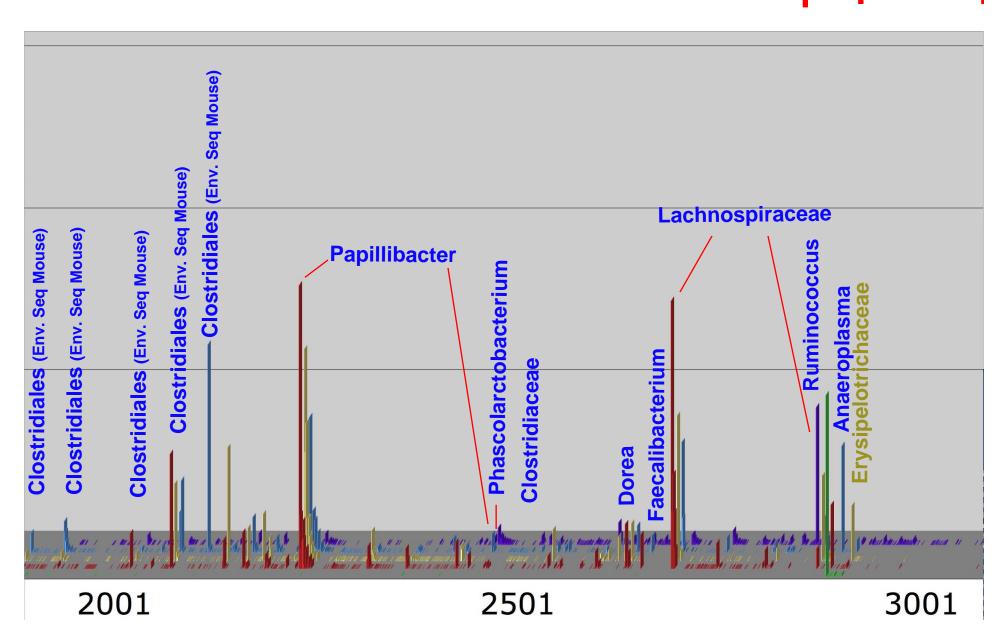
Marine ST 5

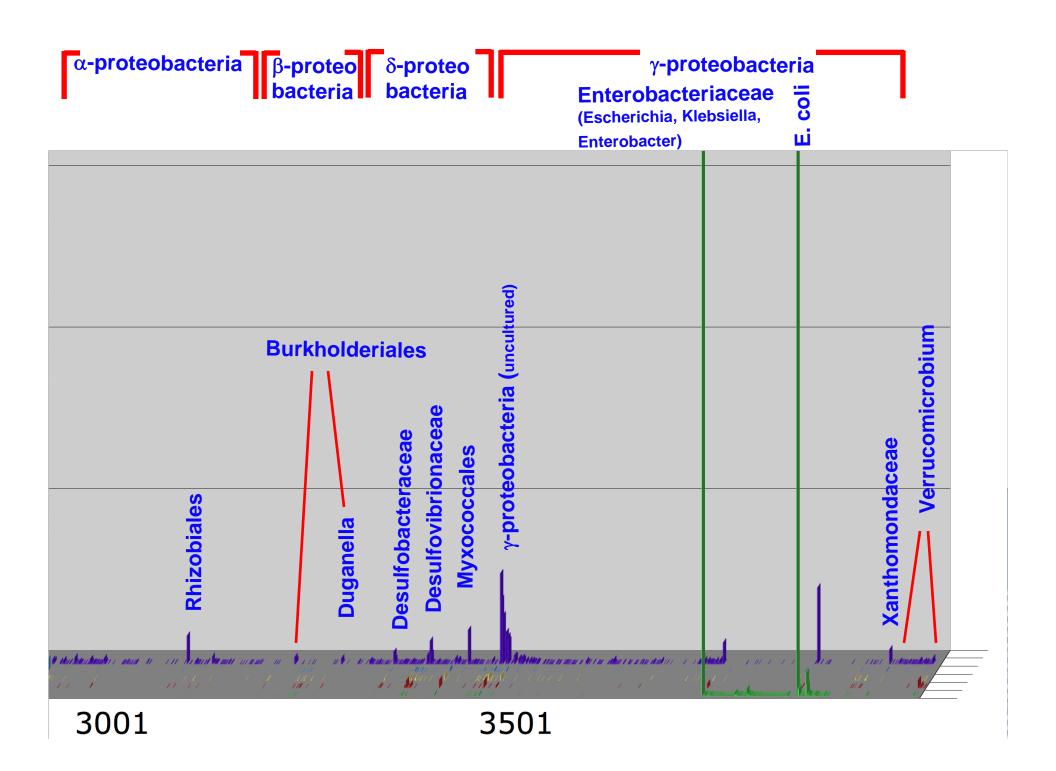
Marine ST 7

QuickTime<sup>™</sup> and a TIFF (LZW) decompressor are needed to see this picture.









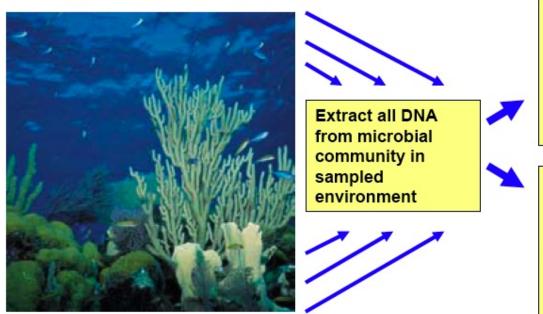
# Why is 454 important?

#### **Benefits:**

- •Can detect and measure minor population members.
- •Provides initial window on total diversity of microbes in the environment.
- •May allow relative quantitative measures of species abundance.
- •Efficiency and high throughput allows intensive sampling of all sites of interest.
- •Offers a tool to finger-print microbial populations for identifying consortia membership and impact on environmental chemistry.
- •Permits microbial population biology studies in natural environments.
- •Makes possible biogeographical studies of microbial populations.
- •Important complement to metagenomic investigations/

## The New Science of Metagenomics Revealing the secrets of Our Microbial Planet

Although we can't see them, microbes are essential for every part of human lifeindeed all life on Earth. The emerging field of metagenomics provides a new way of viewing the microbial world that will not only transform modern microbiology, but also may revolutionize understanding of the entire living world



### DETERMINE WHAT THE GENES ARE (Sequence-based metagenomics)

- Identify genes and metabolic pathways
- Compare to other communities
- and more...

## DETERMINE WHAT THE GENES DO (Function-based metagenomics)

- Screen to identify functions of interest, such as vitamin production or antibiotic resistance
- •Find the genes that code for functions of interest
- •and more...

In order to properly guage expectations, investigators should have some grasp of the population structure of a community prior to carrying out extensive sequencing. In order to provide needed data for the field, it is appropriate to carry out shallow sampling of many environments and deeper, and more thorough sampling in a few environments. In the shallower sampling, for examplie, researchers can compare similar niches to uncover whether they bear the same dominant organism. Deeper sampling can enable scientists to quantify population level heterogeneity, that is, how representative an isolate is of its population. Focusing on the abundant members of a microbial community has a chance of capturing organisms that carry out the dominant processes (such as carbon cycling), but other keystone processes that are carried out by fewer organisms (such as nitrogen fixation) may be missed.

Reconciling Microbial Systematics & Genomics - AAM Colloquium-Berry Buckley and Richard J. Roberts

#### Metagenomics, coding capacity and genomic context

Typical genome size ~2,000,000 bp

Average coverage for Sargaso Sea samples: 200,000,000 bp = 100 genome equivalents

4X Coverage == non-contiguous sequences for ~90% coding capacity 25 genomes

8-15 X Coverage == Assembly into contigs 10-100 kb 6-10 genomes

100 mbp of Metagenomic data from a single microbial community

Coding capacity coverage if a genome represents >4% of organisms

Large contigs if a genome represents 10-15% of the organisms.

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

QuickTime<sup>™</sup> and a TIFF (LZW) decompressor are needed to see this picture.

# Metagenomics is a power method to interrogate function but

Genomic context is important to understand evolutionary history and process

#### **Possible solutions:**

Greater coverage,

**Large Insert Insert Sequencing** 

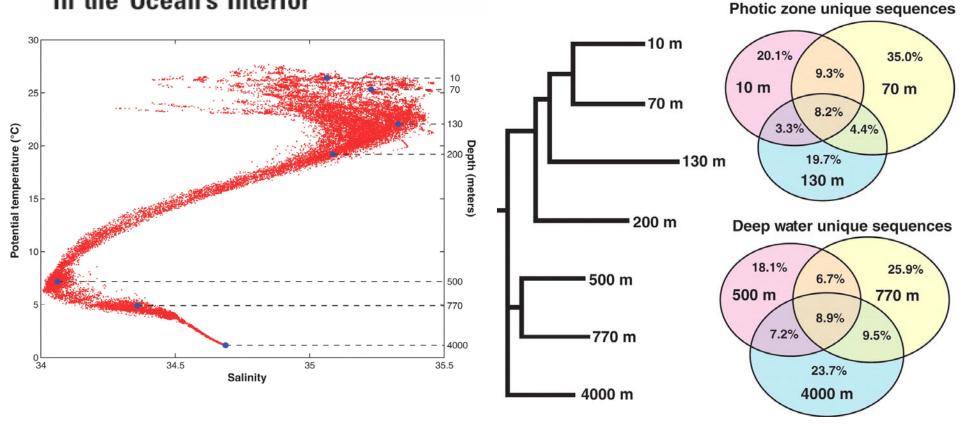
Combinatorial imaging - i.e. simultaneous FISH for up to 500 targets

Single cell genomics / advanced culturing technology

# Microbial Patterns in the Ocean Community Structure and Diversity

27 JANUARY 2006 VOL 311 SCIENCE www.sciencemag.org

#### Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior



# Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time

#### Ramunas Stepanauskas\* and Michael E. Sieracki

Bigelow Laboratory for Ocean Sciences, P.O. Box 475, West Boothbay Harbor, ME 04575-0475 Edited by David M. Karl, University of Hawaii, Honolulu, HI, and approved April 11, 2007 (received for review January 18, 2007)

The identification of predominant microbial taxa with specific metabolic capabilities remains one the biggest challenges in environmental microbiology, because of the limits of current metagenomic and cell culturing methods. We report results from the direct analysis of multiple genes in individual marine bacteria cells, demonstrating the potential for high-throughput metabolic assignment of yet-uncultured taxa. The protocol uses high-speed fluorescence-activated cell sorting, whole genome multiple displacement amplification (MDA), and subsequent PCR screening. A pilot library of 11 single amplified genomes (SAGs) was constructed from Gulf of Maine bacterioplankton as proof of concept. The library consisted of five flavobacteria, one sphingobacterium, four alphaproteobacteria, and one gammaproteobacterium. Most of SAGs. alphaproteobacteria, apart from phylogenetically distant from existing isolates, with 88-97% identity in the 16S rRNA gene sequence. Thus, single-cell MDA provided access to the genomic material of numerically dominant but yet-uncultured taxonomic groups. Two of five flavobacteria in the SAG library contained proteorhodopsin genes, suggesting that flavobacteria are among the major carriers of this photometabolic system. The pufM and nasA

more productive alternative for metabolic mapping of uncultured microorganisms. This strategy has been gaining momentum, with recent implementations of single-cell multiplex PCR in termite gut microbiota by Ottesen et al. (10) and partial genome sequencing of single cells of Prochlorococcus by Zhang et al. (11). Here we show important improvements in single-cell separation and DNA analysis protocols and demonstrate a proof-of-concept metabolic mapping of taxonomically diverse marine bacterioplankton. One of the main challenges for single-cell studies is efficient and contamination-free separation of individual cells from other microorganisms and extracellular DNA in an environmental sample. Prior studies of DNA in individual prokaryote cells used serial sample dilution (11), dilution using microfluidics (10), or micromanipulation (12). Here we used high-speed droplet-based FACS. Compared with alternative methods, FACS offers several critical advantages, including high-throughput rates and the ability to sort targeted plankton groups, based on cell size and fluorescence signals of natural cell components and fluorochromes (13). Furthermore, cell separation by FACS creates microsamples containing the target cell and only 3-10 pl of sample around it (13). This reduces the codeposition of extracellular DNA which in marine waters occurs at

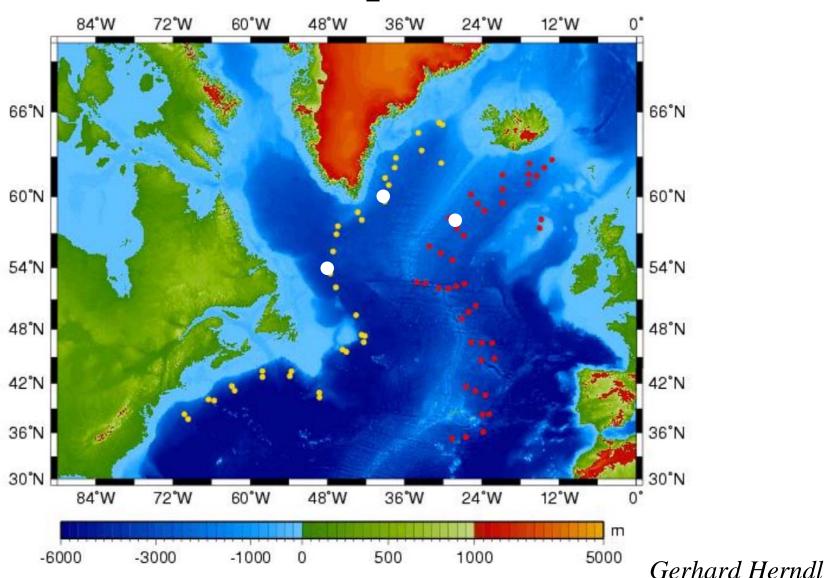
#### **ICoMM Call for Participation**

Unveiling the Ocean's Hidden Majority through 454 Tag Sequencin

W.M. Keck Foundation

The International Census of Marine Microbes (ICoMM) (<u>icomm.mbl.edu</u>) invites proposals from the microbial oceanographic community for V6-tag sequencing of samples from diverse marine habitats and biogeographic provinces of the world's oceans.

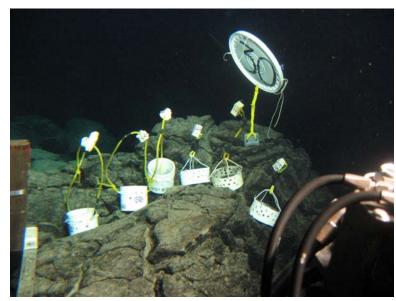
# TRANSAT Samples North Atlantic Deep Water (NADW)



# Hawaii Ocean Time-series: Station ALOHA



## **Basalts and Subseafloor Sediments**



Katrina Edwards

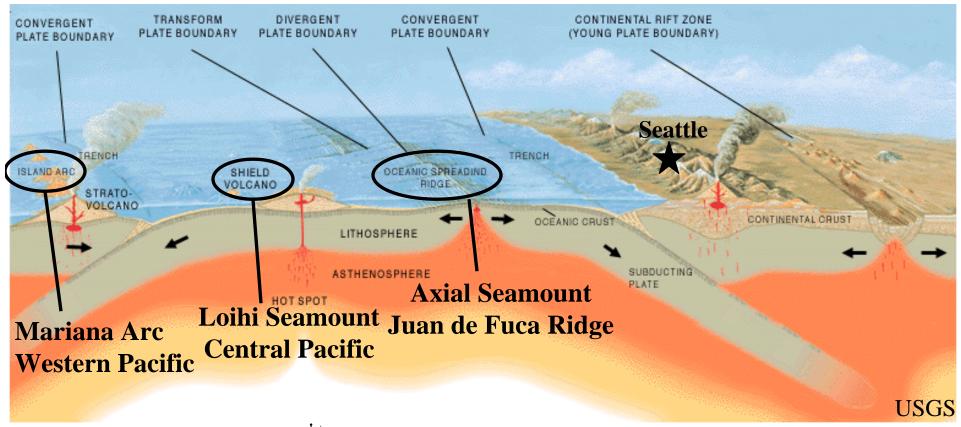


Steve D'Hondt



 $\overline{IODP}$ 

# **Hydrothermal Seamounts of the Pacific**

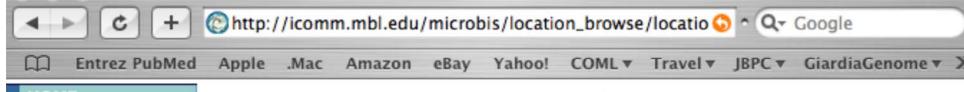








NOAA NeMO Vents



HOME

SEQUENCE SEARCH

INVENTORY BROWSE

TAXONOMIC BROWSE

LOCATION BROWSE

**DOWNLOADS** 

DESIGN DOCUMENTS

XML INFORMATION

LIPIDS DATABASE

**ICoMM** 

OTHER LINKS

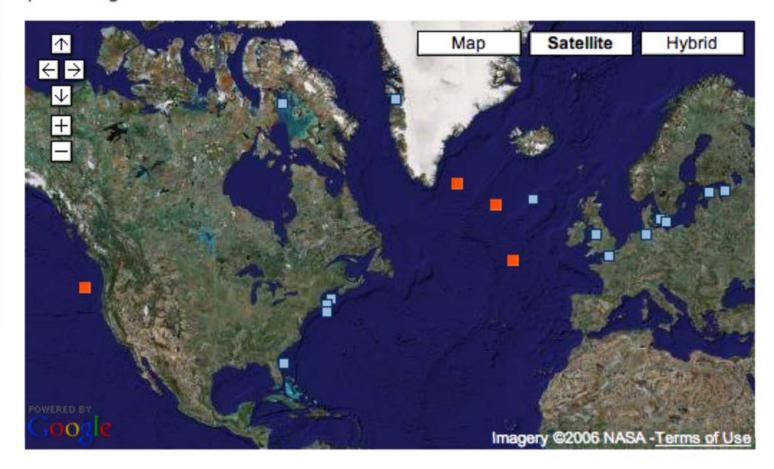
CONTACTS

ABOUT

Madified: Thu Sep 28 12:16:34 2006

#### **MICROBIS**

This is the spatial distribution of sample sites here at MICROBIS. Click on a square to get more information and to drill down.



## Have a good idea?

- ICoMM Call for Participation:

  Unveiling the Ocean's Hidden Majority
  through 454 Tag Sequencing
  - http://icomm.mbl.edu

# Acknowledgements

- Julie Huber, Hilary Morrison, David Mark Welch, Linda Amaral Zettler, Susan Huse, Phil Neal (MBL)
- Gerhard Herndl and Jesus Arrieta (NIOZ)
- Rebecca Gast (WHOI)
- Vince Young (Michigan State University)
- NIH/NSF Centers for Oceans and Human Heal
- NASA Astrobiology Institute
- Alfred P. Sloan Foundation
- W. M. Keck Foundation